Weekly Report

March 11, 2019

1 Work

1. 本周基本完成了ICCV论文的写作以及模型结果的生成和对比。后面还需要不断优化文字和图片等。
2. 工作时长：工作日每天10个小时，周末共16个小时，共66个小时。

1.1 工作进度

Table 1: 工作进度

项目	进度	截止时间
DRGraph	需要对程序做一些修改	2019.4.30
ICCV投稿	写论文中	2019.3.23
unpair 低光照图片增强	准备进行初步实验	

2 Paper Reading

2.1 Holistically-Nested Edge Detection

HED模型基于VGG网络，在每个卷积后面输出一张边缘图，最后通过把不同层次的边缘图融合起来，获得最终的图片。性能好的原因可能是1）使用了VGG网络的参数，2）多层loss共同优化最终结果，3）多层次信息融合

2.2 Towards Open-Set Identity Preserving Face Synthesis

disentangle类型的一篇文章，给定两张图片，一张图片给物体，另一张图片给属性，然后进行合成。

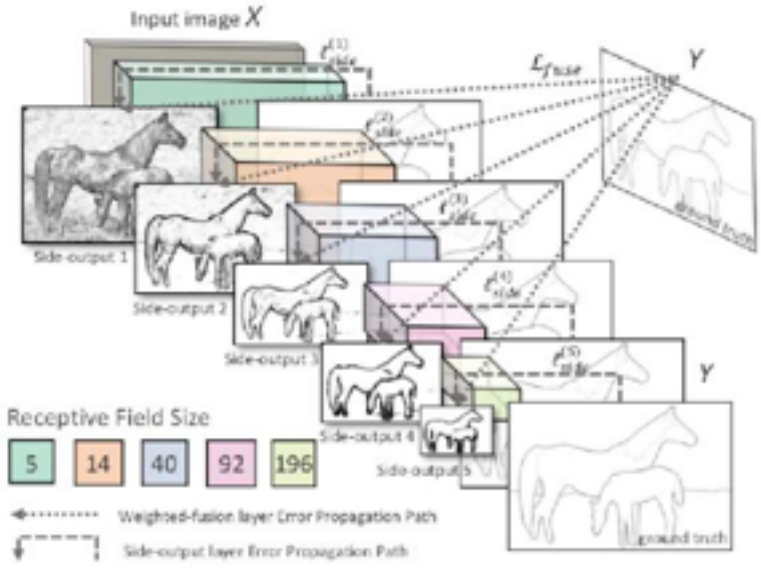


Figure 1: #1

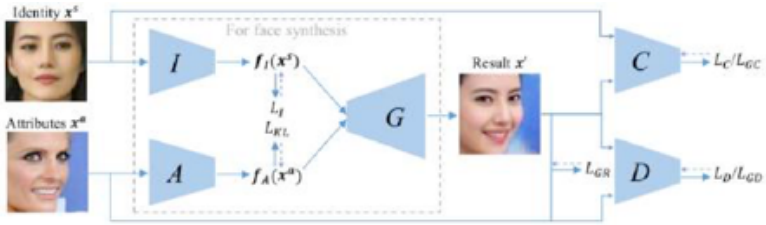


Figure 2: #2

2.3 Multi-Frame Quality Enhancement for Compressed Video

想要增强一个被压缩过的视频的画质，对于某一帧，我们可以使用前后的高质量帧来增强。

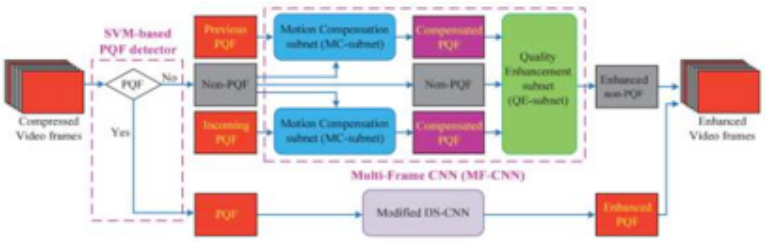


Figure 3: #3

2.4 Learning to Understand Image Blur

本文的任务是给定一张图片，输出各个像素上的模糊度，并且给模糊度分类。

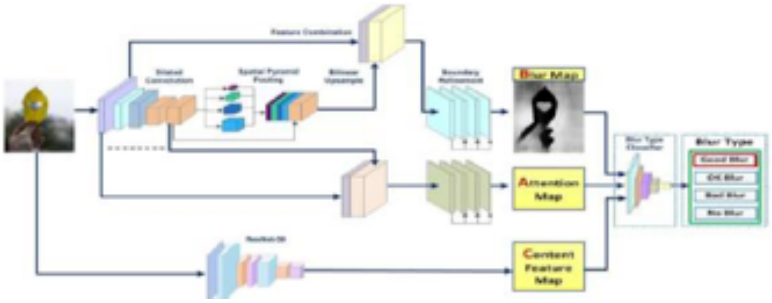


Figure 4: #4

Extreme Low-Light Image Enhancement via Edge Enhanced Multi-Frame Network

Anonymous ICCV submission

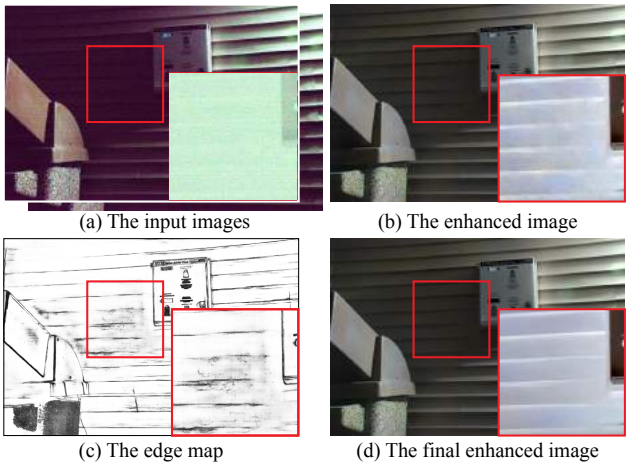
Paper ID 1295

Abstract

Computer vision has been the core of most artificial intelligence applications, whereas images captured in extremely low light conditions may significantly affect the performance of computer vision tasks. Recently, deep learning based low light image enhancement approaches have yielded impressive progress over conventional methods. However, most existing approaches suffer from two main problems when applied to extremely low light images: (1) they suffer from noise and color bias due to challenging scenes in extreme low light condition; (2) l1 loss processes local areas with different local structures equally, which may lead to blurred images and loss of details. In this paper, we propose a new architecture, Edge Enhanced Multi-Frame Network (EEMFN), to enhance extremely low light images while maintaining sharp edges and fine scale details. First, we decrease the color bias and noise by fully exploiting valuable information from two short-exposure low light images. Second, we introduce an edge detection network to predict local structure information. The merge network preserves abundant textures and sharp edges under the guidance of local structures. Experiments on the See-in-the-Dark dataset indicate that our EEMFN achieves state-of-the-art performance in low light image enhancement.

1. Introduction

High-visibility images with clear details are critical to computer vision tasks, e.g., video surveillance and object detection. However, images may lose information in the dark region and receive unexpected noise and color bias when captured in extremely low light condition (see an example in Figure 1(a)). The low quality images may significantly affect the performance of computer vision tasks that rely heavily on the quality of input images [16]. Therefore, in order to recover a high quality image, low light image enhancement techniques is highly desired to remove noise and reveal hidden information from dark regions.



(a) The input images (b) The enhanced image (c) The edge map (d) The final enhanced image

Figure 1. An example of our EEMFN pipeline. The red rectangle represents where the sub image was taken. It can be clearly seen that EEMFN recovers sharp edge efficiently. (a) Two raw images captured by Sony α7S II and scaled by the desired amplification ratio. (b) The enhanced image by multi-frame network. (c) The edge map obtained from the enhanced image. (d) The final enhanced image by combining (b) and (c).

Over the past few decades, various low light image enhancement methods have been proposed to address this problem. Traditional techniques can be divided into two major categories: Histogram-equalization-based methods [14] and Retinex-theory-based methods [13]. These methods recover low pixel values to make them obey a more natural distribution. Recently, convolutional neural network achieves great progress in super resolution, image denoising and low light image enhancement [1, 9]. However, enhancing extreme low light images is still challenging due to the following reasons. First, most existing approaches may suffer from severe noise and color bias. Though a long exposure time can reduce noise and color bias, the image will get blurred by the camera motions and aperture. Alternatively, various methods, such as BM3D [7], are proposed based on human knowledge or known noise like Gaussian noise. However, real scenes are more complex so that the result

of existing models is unsatisfying when applied to extreme low light images. Second, minimizing $l1$ or mean squared error loss to mitigate noise will lead to blurred images and loss of details. They may focus on global information by averaging nearby pixels to achieve quantitatively better performances, resulting in losing sharp edges. To solve the first issue, we fuse multiple short-exposure low light images to decrease noise and color bias. While single-frame image enhancement has achieves remarkable performance, it can further benefit significantly from fully exploiting valuable information accumulated over multiple images. For the second issue, we utilize an edge detection network to predict visually pleasing edge information and preserve local structures. The human visual systems are highly sensitive to edge structures, which often result in better performances, so that it is necessary to guide the network to reconstruct sharp edges and fine image details.

In this paper, we propose a new architecture, Edge Enhanced Multi-frame Network (EEMFN), to enhance extreme low light images while maintaining sharp edges and fine scale details. Generally, our EEMFN consists of three modules: multi-frame network, edge detection network and merge network. Figure 1 illustrates an example of EEMFN pipeline. The multi-frame network decrease noise variance and color bias by fusing two short-exposure low light images. We introduce an edge detection network to predict image edges for fine scale detail restoration. Finally, the merge module yields high quality images by taking advantages of global features from multi-frame module and local features from edge module. We evaluate our method on the See-in-the-Dark dataset and compare ours with previous methods. Qualitatively, our proposed EEMFN produces higher quality images compared to the state-of-the-art methods. For instance, the Peak Signal-to-Noise Ratio (PSNR) of our approach on the Sony set is 29.7 dB, compared to 28.8 dB for the U-net architecture. Our approach improves the Structural SIMilarity (SSIM) on the Fuji set from 0.68 to 0.80. Quantitative results indicate that our EEMFN achieves a more natural result with abundant textures and sharp edge.

In summary, we make following contributions:

- We propose a multi-frame network to decrease noise variance and color bias by combining two short-exposure low light images;
- We introduce a edge detection network for accurately estimating fine scale local structures;
- The experimental results demonstrate that the proposed EEMFN achieves state-of-the-art performance. Furthermore, we conduct an ablation study to demonstrate the effectiveness of each module.

This paper is organized as follows. Section 2 summarizes related work. Section 3 presents our EEMFN model. Experiments are introduced in Section 4, followed by discussions in Section 5. We draw conclusions in Section 6.

2. Related Work

2.1. Low-light Image Enhancement

Low-light images captured in extreme low-light condition will certainly reduce the performance of computer vision algorithms. Thus, various low-light image enhancement approaches have been proposed to recover a high quality image from a low-light image. Traditional approaches can be categorized into two main categories: histogram-based method [14] and retinex-based methods [13]. For example, histogram equalization [6] tries to map the histogram of the whole image as a simple mathematical distribution. However, these methods recover each pixel individually without taking surrounding pixels into consideration. Retinex-based methods [10] first estimate an illumination map according to the retinex theory and then enhance each pixel using the well-constructed illumination map.

Recently, deep learning based methods have achieved significant improvements in image enhancement [5] compared to conventional methods, such as deblurring [1], denoising [9] and low-light image enhancement [4]. LLNet consists of a contrast enhancement module and a denoising module based on the autoencoder architecture. LLCNN [20] applies a special-designed convolutional module to utilize multi-scale feature maps to enhance low-light images. Retinex-Net [22] consists of a Decom-Net for decomposition and an Enhance-Net for illumination adjustment. CAN [5] uses a fully-convolutional network [17] to approximate a variety of processing operators. Chen *et al*. employed a fully-convolutional network based on the U-net architecture [18] for single low-light image enhancement [4]. Although these methods may produce satisfying result sometimes, they may generate blurred images in order to reduce significant noise of extremely low-light images by averaging nearby pixels. In this paper, we propose to intelligently reduce noise and produce a more accurate enhancement by fusing partial information from multiple images.

2.2. Edge Detection

Edge Detection is one of the most fundamental computer vision tasks. Existing methods can be roughly categorized into three groups. The first one usually produces an edge map by designing various filters manually. For instance, Canny [3] introduced Gaussian smoothing in the process of extracting the image gradient. The second category predicts edges using data-driven models according to features of human design. Structured Edges [8] employs random decision forests to learn the structure of edge patches. Third, deep learning learns complex feature representations from raw data and have achieved considerable progress recently.

HED [23] is an end-to-end edge detection model which combines side outputs from multiple scales. Deepedge [2] averages the outputs from a classification branch and a regression branch to produce the final outputs. Deepcontour [19] divides edge data into subclasses and fits each subclass using different model parameters. RCF [15] uses richer features from all the convolution layers to perform an image-to-image prediction task in real-time. Liu *et al.* proposed diverse deep supervision which minimizes different loss functions for high-level and low-level feature learning. Given that the human visual system is highly sensitive to the edges, preserving edge information is crucial to the performance of image reconstruction task. SREdgeNet [11] employs a novel edge detection network to provide edge information as a guide to yield better and more realistic super-resolution images. Inspired by previous works, we propose to reconstruct a high quality image with abundant textures and rich local structure using edge information for low-light image enhancement task.

3. Method

Our goal is to enhance extremely low-light images with noise. We propose a novel architecture, Edge Enhanced Multi-frame Network (EEMFN), that is trained for enhancing low-light images while denoising real noises, reducing color bias and maintaining sharp edges. As illustrated in Figure 2, our proposed EEMFN model consists of three modules: multi-frame network, edge detection network and merge network. A brief description of each network is listed as follows:

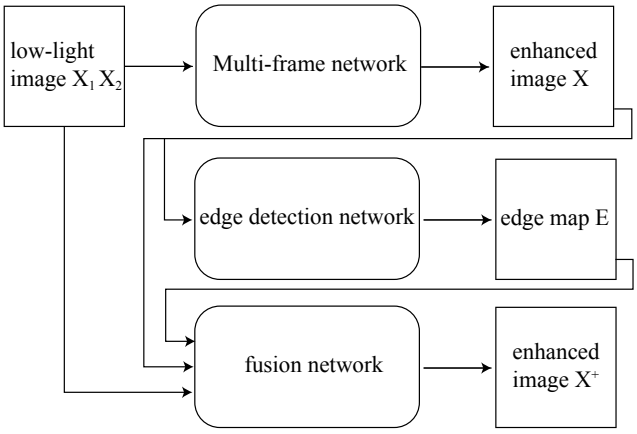


Figure 2. Demonstration of our EEMFN pipeline for low-light image enhancement. The proposed EENet consists of three networks: multi-frame network, edge detection network and fusion network.

Multi-frame Network. Given raw images I_1^{raw} and $I_2^{raw} \in R^{H\times W\times 1.}$, the multi-frame network is introduced to combine two low-light images in to a single image which holds the meaningful information from both images.

$$I = MFNet_\theta(I_1^{raw}, I_2^{raw}), \tag{1}$$

where $MFNet$ denotes the function of multi-frame network and $I \in R^{H\times W\times 3}$ is the output initial image. As shown in Figure **??**, we implement $MFNet$ by exchanging convolutional features between two images when we process each frame with the same U-net [18].

Edge Detection Network. The second module is an edge detection network which predicts an edge map from the output of the multi-frame network. Then the edge information is utilized to guide the reconstruction of high quality images with sharp edges.

$$E = EdgeNet(I), \tag{2}$$

where $I_i^{edge} \in R^{H\times W\times 1}$ is the edge map of the image I and $EdgeNet$ is the function of edge detection network.

Merge Network. The merge network predicts the final image by fusing low level and high level information from the output of the multi-frame network and edge detection network.

$$I^+ = MergeNet(I_1^{raw}, I_2^{raw}, I, E), \tag{3}$$

where I^+ is the final enhanced image and the $MergeNet$ denotes the function of merge network.

3.1. Multi-frame Network

Figure 3 illustrates the architecture of multi-frame network, which is fed with two noisy low-light images and produces an enhanced image. The key idea is that each individual frame is processed separately by the same network to extract global representations and then several exchange blocks are employed to exchange their local representations for a high quality image.

First, each branch process a input image using the same U-Net with skip connections to aid the reconstruction of details at different scales.Second, each **exchange block** takes two image features $F_1, F_2 \in \mathbb{R}^{H\times W\times C}$ from two branches as the input and performs max and average operations to extract local features for sharing. This yields a feature tensor of size $H \times W \times 2C$.

$$F_{max}(i,j,k) = \max(F_1(i,j,k), F_2(i,j,k)), \tag{4}$$
$$F_{avg}(i,j,k) = (F_1(i,j,k) + F_2(i,j,k))/2, \tag{5}$$
$$F = [F_{max}, F_{avg}], \tag{6}$$

where $[\cdot]$ denotes the concatenation operation,F_1 and F_2 is the correspondingly sized feature from two branches. We transform the feature tensor into the input feature space.

$$O = W_{ex}F \tag{7}$$

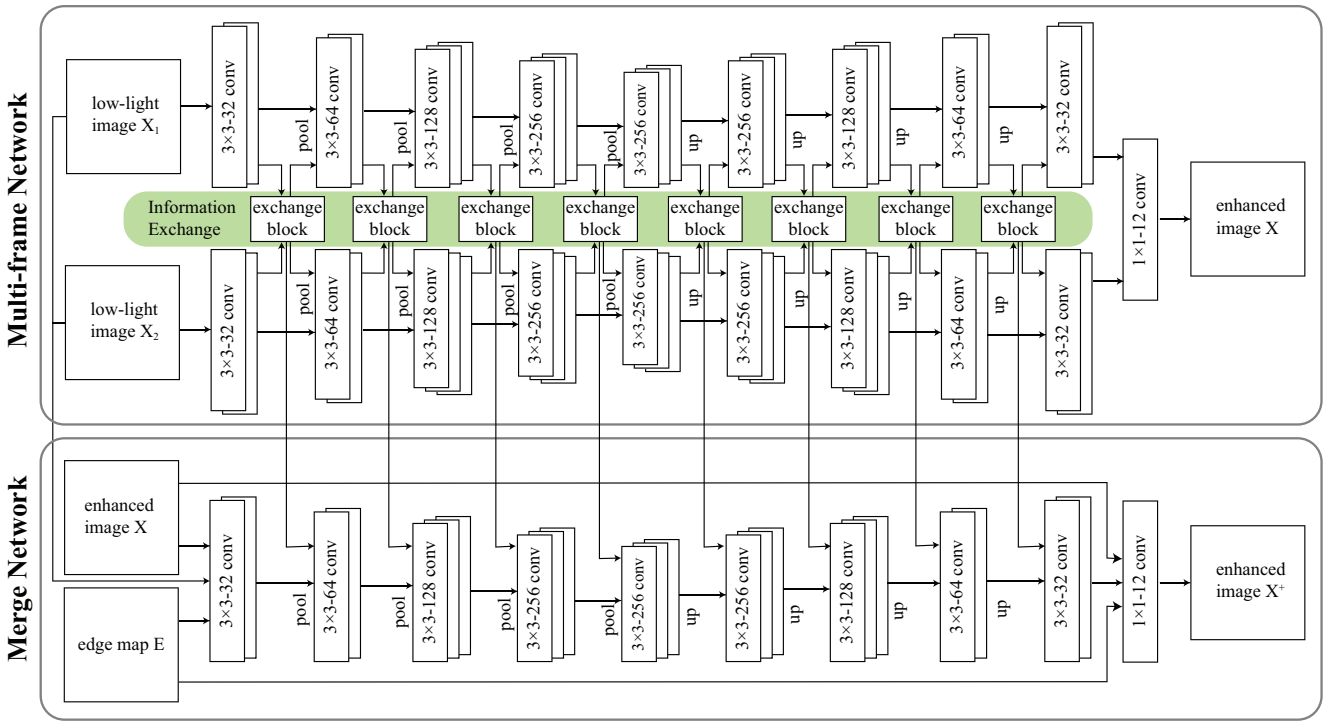


Figure 3. The architecture of our multi-frame network. Given two input images, we extract global features using the same U-net [18] separately and exchange local information between two branches.

where $W_{ex} \in \mathbb{R}^{C\times 2C}$ are the learned weight matric, which is implemented by a 1×1 convolution. Then, the local features which contain lots of useful fine details are fed back into each branch. Finally, the output features of two branches are concatenated together and fed into a $1\times1 Conv$ layer for a joint estimate of a clean image. The objective function of the multi-frame module is define as l_1 loss between the output image and the ground truth:

$$L_{multi} = \|I - I^{gt}\|_1. \tag{8}$$

3.2. Edge Detection Network

To train the edge detection network, we generate a set of input-output pairs where the input is the ground truth image and the output is the corresponding edge map computed by using Canny edge detector [3]. We employ RCF network [15] to predict edges. The RCF network consists of five stages, each of which makes use of all the activation of convolution layers to perform the pixel-wise prediction $(E_1, E_2, E_3, E_4, E_5)$. Finally, by carefully combining hierarchical CNN features of all the stage, the edge detection module can obtain more accurate edge map E.

Considering the fact that the distribution of edge/non-edge pixels is heavily unbalanced, we compute a weighted cross entropy loss at each pixel with respect to pixel label. We employ two class-balancing weights α and β to offset such unbalance. The image loss between predicted edge map $E = (e_j, j = 1, ..., |E|), e_j = \{0, 1\}$ and ground truth $E_{gt} = (e_j, j = 1, ..., |E_{gt}|), e_j = \{0, 1\}$ is defined as:

$$\begin{aligned} l_{edge}(E_i, E_{gt}) = &- \alpha \sum_{j\in E^+_{gt}} \log \Pr(e_j = 1|I, i) \\ &- \beta \sum_{j\in E^-_{gt}} \log(1 - \Pr(e_j = 1|I, i)), \end{aligned} \tag{9}$$

$$\alpha = \frac{|E^-_{gt}|}{|E^+_{gt}| + |E^-_{gt}|}, \beta = \frac{|E^+_{gt}|}{|E^+_{gt}| + |E^-_{gt}|}, \tag{10}$$

where $|E^+|$ and $|E^-|$ denote the size of the edge and non-edge ground truth label sets and $\Pr(\bar{e}_j = 1|I, i)$ is the activation value of pixel j at stage i.

Then, a 1×1 convolutional fusion layer is employed to combine the output edge map from all stages:

$$E = Conv_{1\times1}(E_1, ..., E_K), \tag{11}$$

The objective function of the edge detection module is computed by aggregating the loss function from different stages and fusion layer:

$$L_{edge} = \sum_{i=1}^{K} l_{edge}(E_i, E_{gt}) + l_{edge}(E, E_{gt}), \tag{12}$$

where $K(= 5)$ is the number of stages.

3.3. Merge Network

The merge network takes the input images, enhanced image and edge map as inputs and integrates these images by taking advantages of global features from multi-frame network and local features from edge network. The network architecture for fusion network is illustrated in Figure 3. A U-net [18] is adopted to produce the final enhanced image I^+. To efficiently utilize pre-computed information by integrating enhanced images and edge maps, we add two edge skip connections to connect the edge map E and enhanced image I for residual learning.

In addition, we propose an edge-preserving loss to guide the network to focus on the discontinuities in the image gradient. The edge-preserving loss is defined as the distance between the edge map of the final output and the corresponding edge map ground truth:

$$l_{EP}(I^+, E) = \|EdgeNet(I^+) - E\|_1. \quad (13)$$

The objective function of the fusion module is defined as:

$$L_{fusion} = l_1(I, I^{gt}) + \lambda l_{EP}(I^+, E), \quad (14)$$

3.4. Implementation and Training Details

Following [4], we preprocess all raw images by subtracting the balck level and scaling the data by a desired amplification ratio, which is the exposure difference between the input and reference image. The raw data of Sony set is packed into 4 channels for every 2×2 blocks. The raw data of Fuji set, which is arranged in 6×6 blocks, is packed into 2×2 blocks with 9 channels. Our model takes the processed Sony data with smaller spatial resolution as input, outputs a 12-channel (27-channel for Fuji set) with the same spatial resolution, and then rearranges data from depth into blocks of spatial space to get a full-resolution image. For training set, we randomly select two frames in a sequence and their corresponding long-exposure image as a input-output pair to train multi-frame module. For test set, we select the first and second frames in a sequence in the test set to predict the ground truth image.

We implemented our proposed EEMFN with Tensorflow framework on a single computer with two Nvidia GTX 1080 Ti. We trained EEMFN using ADAM [12] optimizer with an initial learning rate of 10^{-4}. We decrease the learning rate to $5 * 10^{-5}$ after 2500 epochs and 10^{-5} after 3500 epochs. We train all the network with 5000 epochs. In experiments, we found that the whole network trained from scratch converges slowly and achieves lower performance. To address this issue, we train each task separately. Thanks to this training strategy, the EEMFN converges much faster than the one started from scratch and achieves a better performance.

4. Experiments

In this section, we evaluate the EEMFN model quantitatively and qualitatively on the See-in-the-Dark dataset, compared with the state-of-the-art methods.

Sony set	Exposure time	# Training	# Test	# Val
x100	0.1s	587	27	11
x250	0.04s	457	27	11
x300	0.1s	753	20	9
x300	0.033s	67	14	5
Fuji set	**Exposure time**	**# Training**	**# Test**	**# Val**
x100	0.1s	804	38	17
x250	0.04s	427	27	13
x300	0.033s	422	26	8

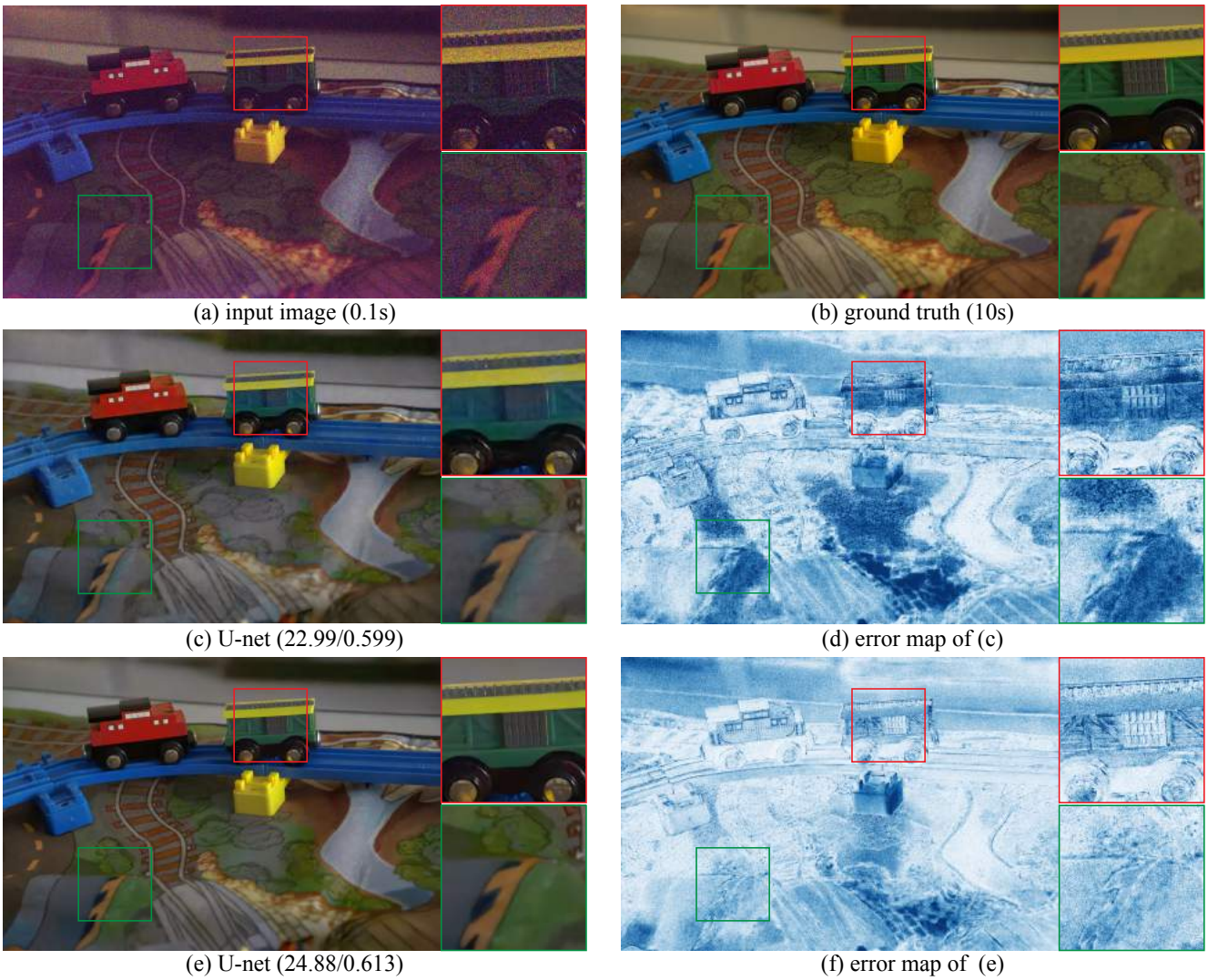
Table 1. Statistics of short-exposure images in See-in-the-Dark dataset which consists of two subsets. The first column indicates the ratio of exposure times between short-exposure and long-exposure images. Note that multiple short-exposure images can correspond to the same long-exposure image.

Dataset. To demonstrate the capability of our proposed method for low-light image enhancement, we conducted experiments on the See-in-the-Dark dataset [4]. The See-in-the-Dark dataset consists of two image set: Sony set and Fuji set. The exposure for the input images was set between 1/30 and 1/10 seconds. The corresponding reference images (ground truth) were captured with 100 to 300 times longer exposure. Multiple short-exposure images can correspond to the same long-exposure image. The statistics of the See-in-the-Dark dataset are summarized in Table 1. The Sony set captured by Sony $\alpha 7S$ includes 1988 raw short-exposure images and 228 corresponding long-exposure reference image. The Fuji set captured using Fujifilm X-T2 contains 1782 raw short-exposure images and 193 long-exposure reference image. The resolution is 4240×2832 for Sony and 6000×4000 for the Fuji images.

Evaluation Metric. We employ the Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM) [21] to evaluate the performance.

	Sony set		Fuji set	
Model	PSNR	SSIM	PSNR	SSIM
CAN [5]	27.40	0.792	25.71	0.710
Chen *et al.* [4]	28.88	0.787	26.61	0.680
baseline	28.98?	0.790?	27.13?	0.700?
EEMFN	**29.78**	**0.802**	**28.02**	**0.730**

Table 2. Quantitative evaluation of low-light image enhancement algorithms in terms of PSNR and SSIM. The best results are highlighted in bold.



(a) input image (0.1s) (b) ground truth (10s)
(c) U-net (22.99/0.599) (d) error map of (c)
(e) U-net (24.88/0.613) (f) error map of (e)

Figure 5. Qualitative results for extreme low-light image enhancement by U-net and our EEMFN on images from the Fuji set.

(a) unet (b) baseline + EB (a) baseline + EB + Edge

Figure 6. The controlled experiment results of ablation study.

4.2. Qualitative Evaluation

Figure 4 and Figure 5 shows some representative results for visual comparison. We show the input image, the output result of U-net and EEMFN, the ground truth and the error map between the output and the ground truth. The error maps show per-pixel error, measured by l_1 distance in RGB space. As we can see, these input images essentially suffer from noise and color bias, which greatly affects the efficiency of our vision. The input images captured in extreme

low-light condition contains more severe noise. We cannot observe the book name (red rectangle) or the general edges of the chair (green rectangle). Although U-net handles the noise and color bias effectively, the result image is blurry and the object edge is difficult to recognize. For Sony set, we are able to see the book name and chair edges clearly, which indicates that the U-net cannot well reconstructed local edge structures. By checking the details, it is clear that our method achieves better visual effects, including small color bias, less noise and richer details. From figure 4(c), we make the following observations: 1) the error map indicates that a smaller error achieved by our EEMFN model; 2) The EEMFN model is able to better refine the edges in the enhance image. Under the guidance of edge information, the error map is more smooth in each object surface. 3) The proposed method is able to better preserve the local structural details. For instance, the texture of the meeting table are better recovered by our model. For Fuji set, one may observe from Figure 5 that our reconstruct high-quality images with sharp edges more effectively than U-net. Specifically, the severely distorted content, e.g., words and textures, can be well restored by leveraging edge information. However, U-net can hardly reduce such distortion, resulting in fuzzy words. In summary, our proposed EEMFN model can recover sharper and clear edges such as the edges of curves in fonts, the texture and structure of objects and the boundary between objects, resulting in much better visual quality.

4.3. Ablation Studies

For a comprehensive understanding of our model, we conduct ablation experiments to demonstrate the improvements obtained by each component. To this end, we perform the following four experiments:

baseline, which has the same network structure as Chen *et al*. [4] and concatenates two input low-light images before feeding them into the network.

MFN\EB, which removes the exchange block from multi-frame network, such that the information of two U-net branch are fused at the last layer.

MFN\avg, which removes the average operation.

MFN\max, which removes the max operation.

MFN, which is the first step of our EEMFN and produce an initial image for further enhancement.

EEMFN, which predicts the final enhanced image by taking edge information into consideration.

We evaluate EEMFN model and its variants on the See-In-The-Dark dataset. Table 3 show the evaluation results of six experiments. We can see that MFN\EB achieves a slightly better performance than baseline. Because MFN\EB process each image individually which increase the width (number of channels) of the network. Furthermore, MFN\avg and MFN\max performs better than MFN\EB, because they transmit information between two branches to make full use of the partial information from each input image. MFN combines two types of pooling operation to further improve the performance. Moreover, the significant improvement provided by our proposed EEMFN over MFN demonstrates the effectiveness of our design, which merges partial information from different images under the guidance of local edge structures.

We also illustrate a visual comparison among baseline, MFN and EEMFN in Figure 6. Baseline suffers from loss of color when recover the color using low-light images. Hence, MFN consists of exchange blocks to make full use of the partial information by concatenating high-quality image features back into each branch. However, the result of MFN still suffers from severely distorted content, because MFN may average pixels of two objects without the guidance of edge information. Our proposed EEMFN is able to reconstruct high-quality images with abundant textures, sharp edges and smooth surface.

In summary, the experimental results demonstrate the effectiveness of multi-frame network with exchange block and edge detection network, which leads to consistent improvement.

	Sony		Fuji	
Architecture	PSNR	SSIM	PSNR	SSIM
Baseline	29.31	0.793		
MFN\EB	29.30	0.793		
MFN\avg	29.49	0.793	27.67	0.722
MFN\max	29.44	0.794		
MFN	29.55	0.795		
EEMFN	**29.78**	**0.802**	**28.02**	**0.730**

Table 3. Quantitative performance of different architectures evaluated on the Sony set and Fuji set. The best performance of each column is highlighted in bold.

5. Conclusions

In this work, we propose a novel deep learning approach, EEMFN, for low-light image enhancement. Instead of increasing exposure time, we decrease noise variance and color bias by fusing multiple short-exposure low-light images. Also, we introduce an edge detection network to reconstruct fine scale details. By merging multiple low-light images and their edge information, EEMFN takes advantages of global and local feature and yields high quality images. Our experimental results have shown that our model could outperform against the state-of-the-art in terms of PSNR and SSIM. The enhanced images generated by the proposed method have good visual quality with sharp edges. In the future, we plan to develop EEMFN with more powerful and faster architectures for real-time processing and apply the model to harder enhancement tasks (e.g. low-light video enhancement).